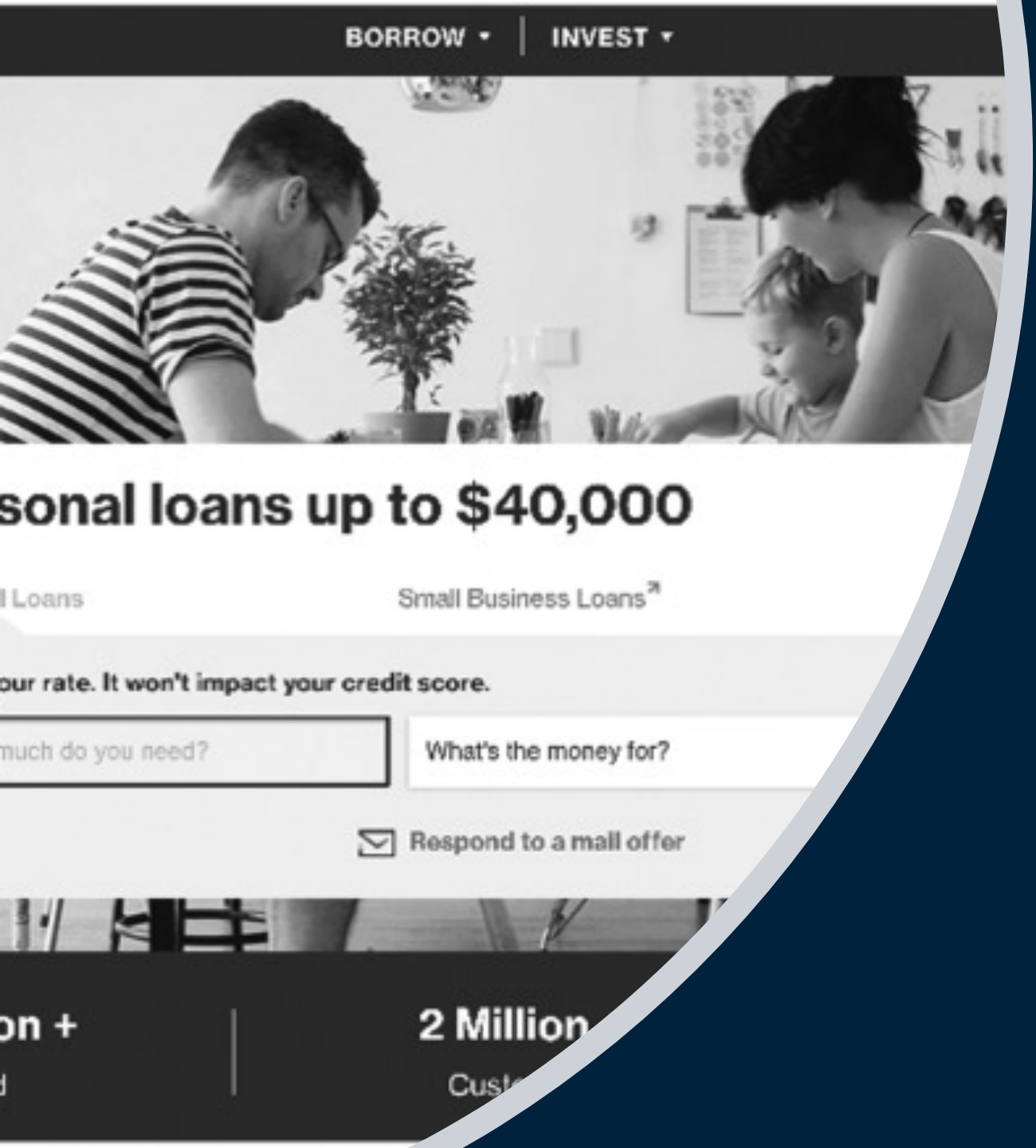


Course Recap

My Final Project: Lending Club

Lecture Plan

- Final Project – Report and Presentation
- My final project – Lending Club

Final Project Report

The final project will your opportunity to present what you have learned in this course and receive constructive feedback. It should be comprehensive, thorough and concise, ***no longer than 10 of ANSI letter pages***.

The report should include the following subsections:

- Introduction
- Key objectives
- Data Source
- Data Structure
- Learning Model
- Calibration
- Results
- Conclusions
- Bibliography

Final Project Presentation

- Final project presentation on Monday May 13th
- Presentations will be done in a job interview format

- **Part 1:** 3-5 minutes monolog response to the question *"tell me about the last project you worked on"*
  - Short overview of the project roughly corresponding to the introduction
    - Do not plan to use slides or any other visualization in the overview
    - You can use a whiteboard for drawing

Final Project Presentation

- **Part 2:**
  - Job interview-style questions about the project
    - Will cover model, methodologies etc.
    - May addressing any part of the project
    - May require write code on the whiteboard
  - It is good practice to try to anticipate some of questions you may be asked and practice answering
  - Teams should present the project together
    - it will be my choice to whom to address a question, so both members must be prepared.
  - I strongly advise you to practice ahead of the presentation with your peers

LendingClub

An example of Final Project Report and presentation

Introduction

- Opening statement and short overview of the topic
- What is interesting with this project
- Clear specification of the research question
  - data science problem
  - business problem
- Why answering the data science problem helps address the business problem.

Overview

- The lending club project explores the opportunities for applying data science methods toward improving investment decisions in a peer to peer lending platform
- LendingClub is a peer-to-peer lending platforms operating since and 2006, connecting borrowers with investor-lenders through an online platforms.
- Online peer-to-peer platforms enable borrowers and lenders to bypass the high operational costs of the traditional banking system to offer borrowers lower interest rates and investors higher returns.

What is interesting?

What I find most interesting in this project is the ability to use data science methodology not only to improve efficiency of the system as an afterthought, but to improve market design by incorporating algorithms into the mechanism to improve the pricing methodology thus gaining better allocations and greater overall efficiency

Research Question

- Data Science question:

  Can the likelihood of a lender to return a loan be forecasted by the a set of features tied to the financial and behavioral characteristics of the borrower?

- Business Question:

  Can a forecasting model for loan return improve investment decisions on this platform by identifying borrowers that are more likely to be profitable investments

Why the Data Science problem helps address the business problem

- The data science problems enables us to address the issue to what effectiveness can the likelihood of a borrower returning a loan be forecasted be estimated at time of borrowing

- This model will enable the lender to make a business decision on whether or not to make a loan to a particular individual and on the pricing decision on what return we are to ask for the risk lender is taking

Key objectives

- Key objectives of the study
- What questions you are addressing?
- What would qualify as a satisfiable answer
- How you would evaluate the answer
- Objectives should address both the data science problem and the business problem

Data

- Sources of data
  - where was data obtained
  - for second hand sources such as Kaggle, what was the source of the original data?
- How big is the data
  - physical storage space (1.8G bytes)
  - coverage (what period is covered, how many individuals are included, how many games etc.)

Data

- The Lending Club platform collects a variety of personal and financial information from borrowers and assigned grades and sub-grades to each loan application for potential investor-lenders to review

- The data is saved on the platform's database and after proper ammonization can be downloaded as csv files by year and quarter
- For development of the model I used data from 2016, 2017 and 2018

- Each file was about 150M large, so each year required 600M and overall 1.8G

Data Structure

- Data ingestion (csv file, API download, h5, SQL query, etc.)
- Data structure used to hold the data in memory (python list or dictionary, NumPy matrix, pandas Series or DataFrame)
- in case of a DataFrame table, specify the columns of the table
- If case casting was applied on what columns or entries and why?
- This section should include snippets of the code used to ingest the data structures used.

Data Cleaning and Validation

- What procedures had you applied for validating the data?
- What fraction of the data was disqualified?
- What procedures did you use to clean the data?

Data Cleaning and Validation

In [22]:

```
for i in float_cols:
    final_data[i] = final_data[i].astype(float)

def clean_perc(x):
    if pd.isnull(x):
        return np.nan
    else:
        return float(x.strip()[:-1])
for i in perc_cols:
    final_data[i] = final_data[i].apply( clean_perc )

def clean_date(x):
    if pd.isnull(x):
        return None
    else:
        return datetime.datetime.strptime( x, "%b-%Y").date()
for i in date_cols:
    final_data[i] = final_data[i].apply( clean_date )

for i in cat_cols:
    final_data.loc[final_data[i].isnull(), i] = None
```

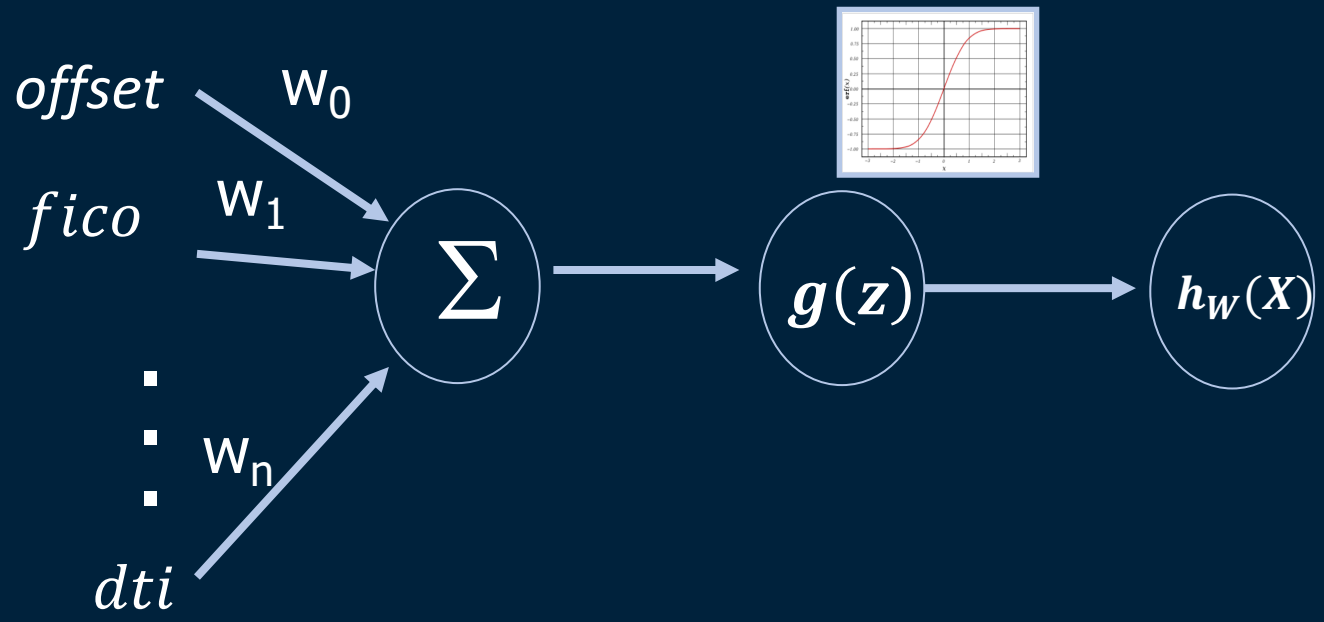
Learning Model

- Learning model used
- Theoretical discussion of the model
  - why it was appropriate for the specific problem
- Libraries were used to implement the model,
- Annotated snippets of code

Learning Model

Theoretical underpinnings

The learning model is based on Logistic Regression



Learning Model

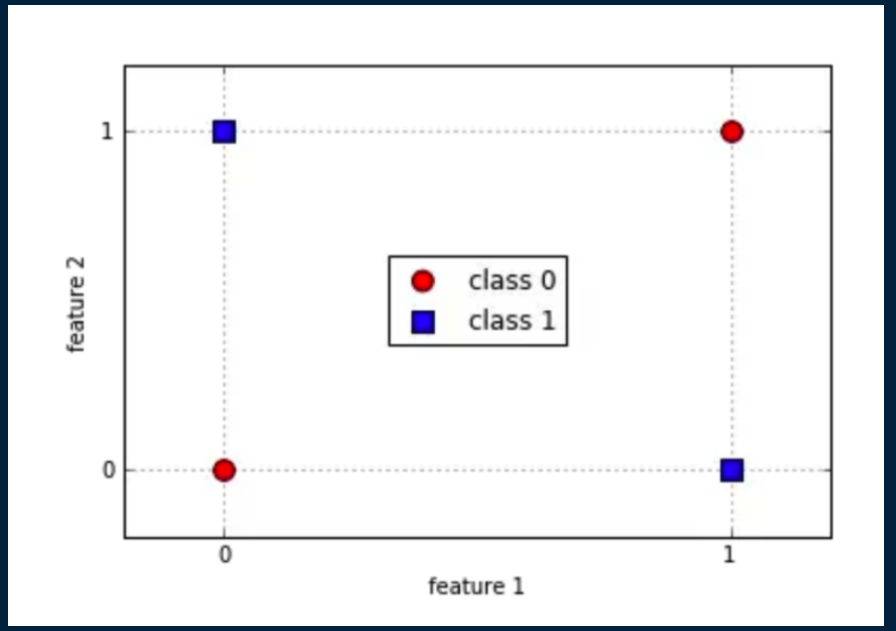
Why is logistic regression an appropriate model?

- Logistic Regression is a regression model that can be used as either a classifier or a regressor.
- A classifier can be used to forecast which borrowers will default on their loan and which will not. We can also use several classifiers to estimate by categories, which will return a loan in full, which will return part of the loan and default and which will default early on
- A regressor can be used to estimate the probability of a default

Learning Model

Weaknesses of the model

- Relies on proper presentation, correlated features, and features that are not correlated with the output tend to degrade performance
- Does not perform well on problems that are inherently non linear



Learning Model

Libraries

In [ ]:
```
from sklearn.linear_model import LogisticRegression

final_features = [i for i in discrete_features + continuous_features
                  if i not in ["grade", "int_rate", "installment"]]

data_dict = prepare_data(feature_subset = final_features)

logistic = LogisticRegression(penalty = 'l2', C=np.inf, solver='lbfgs')

logistic = fit_classification(logistic,data_dict,
                              model_name = 'penalized logistic l2')
```

Fitting, Validation and Evaluation

- What is the parametric setting of the?
- How were the parameters chosen for the model?
- How does this setting protect from overfitting?
- What is the validation of the model.
- How is the overall performance of the model assessed and does this performance fall in line with the objectives initially set for this project?

Learning Model

Libraries

In [ ]:
```
def fit_classification(model,data_dict,cv_parameters={},model_name=None):

    # --------------------------
    #   Step 1 - Load the data
    # --------------------------
    X_train = data_dict['X_train']
    y_train = data_dict['y_train']

    X_test = data_dict['X_test']
    y_test = data_dict['y_test']

    filter_train = data_dict['train_set']
```

Learning Model

Libraries

In [ ]:

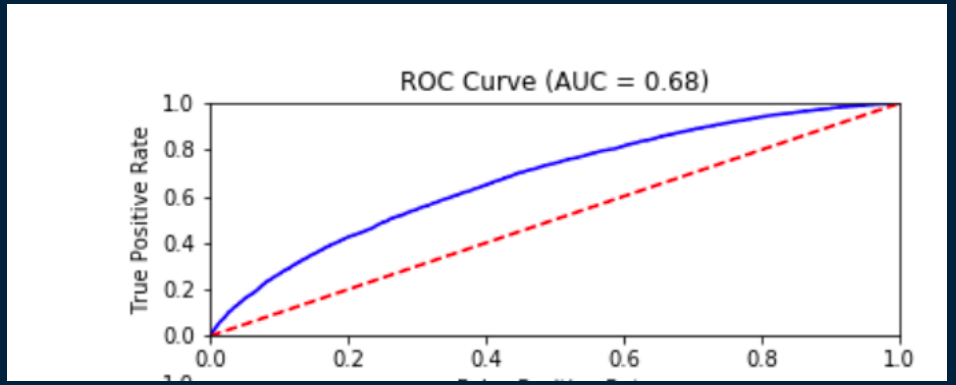
```
# -------------------------
#   Step 2 - Fit the model
# -------------------------
cv_model = GridSearchCV(model, cv_parameters)
start_time = time.time()
cv_model.fit(X_train, y_train)
end_time = time.time()
best_model = cv_model.best_estimator_
```

Learning Model

Libraries

In [ ]:
```
# -------------------------------
#   Step 3 - Evaluate the model
# -------------------------------

    y_pred_probs = best_model.predict_proba(X_test)[:,1]
    fpr, tpr, thresholds = roc_curve(y_test, y_pred_probs)
```

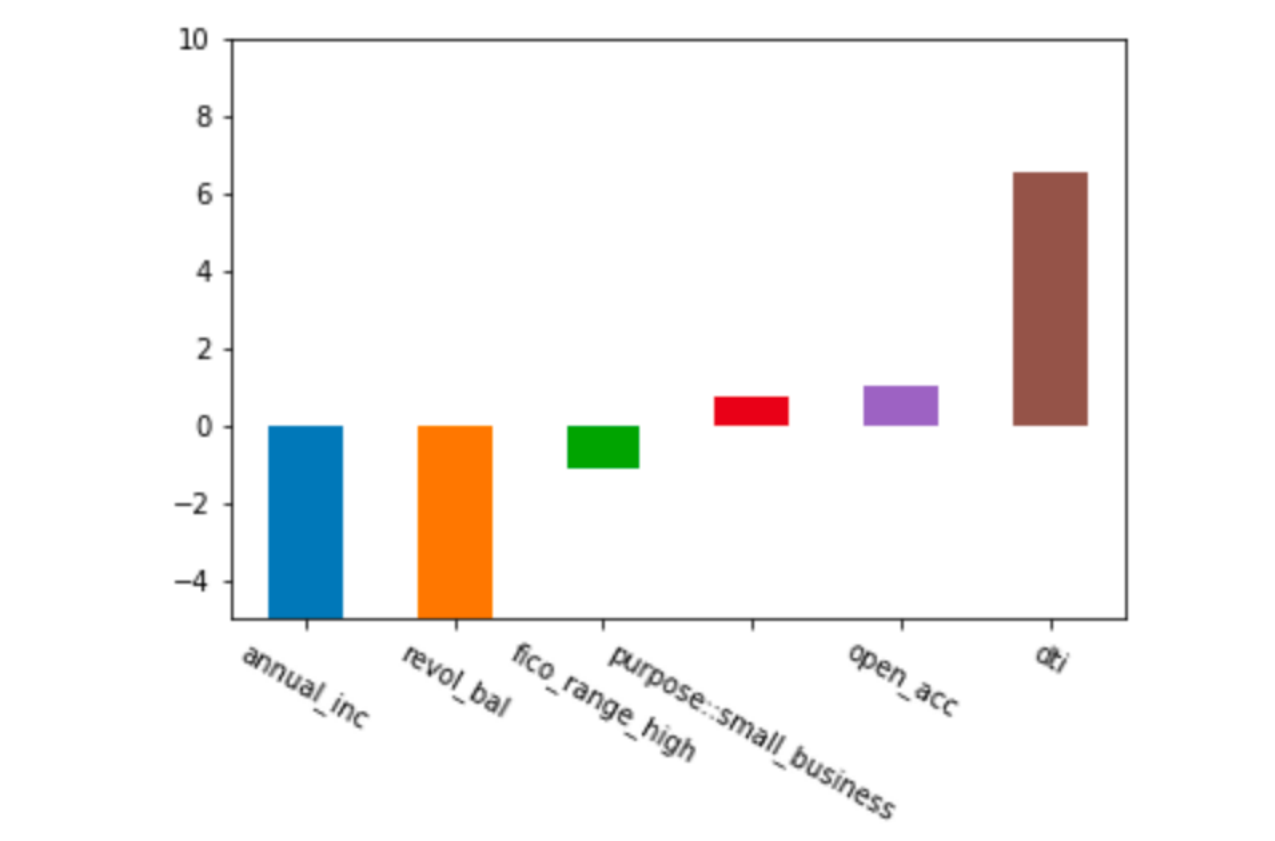


Results

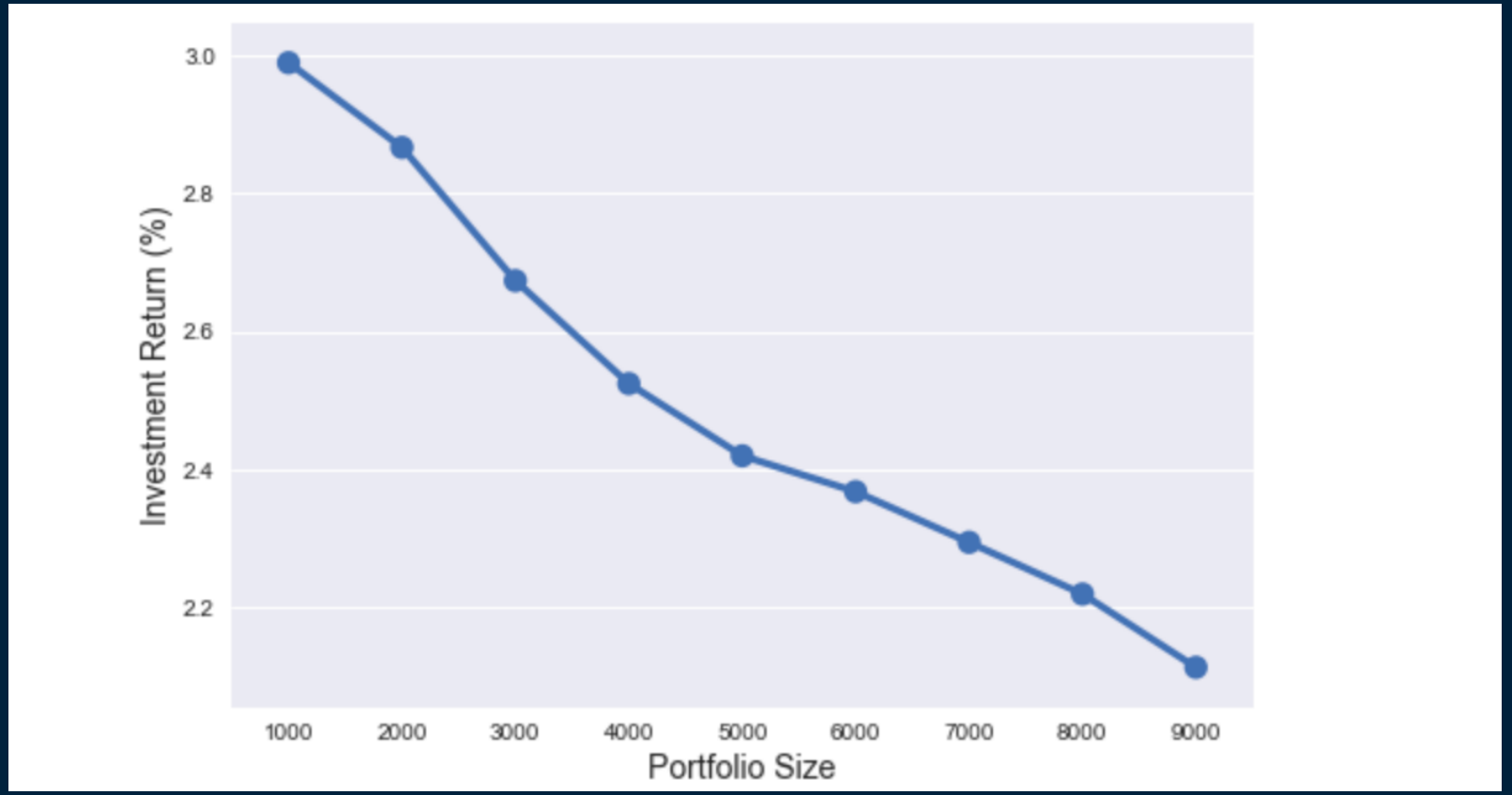
- This section should present the results (tables, visualizations, charts, graphs etc.)
-
- Make sure the results show the full diversity of the data used and clearly present the outcomes.

- Two or three key visualizations that support the main conclusions

Results


10
8
6
4
2
0
−2
−4
annual_inc
revol_bal
fico_range_high
purpose::small_business
open_acc
dti

Results


3.0
2.8
2.6
2.4
2.2
Investment Return (%)
1000
2000
3000
4000
5000
6000
7000
8000
9000
Portfolio Size

Conclusion

- Main conclusions should be clearly stated
- Conclusions to the data science problem
- Conclusion of the business problem

Conclusion

Conclusions to the data science problem:

With the features provided by LendinClub default by borrowers can be forecasted with

Accuracy 0.74485
Precision 0.6998
Recall 0.7449

Conclusion

Conclusion of the business problem

An investment strategy based on logistic regression model for forecasting default using a variety of features can improve the returns over a baseline of random lending as well as a logistic regression based model using only fico score